Purpose: The rpubchem and rcdk are used in R console in purpose of generating a heatmap to show the similarity of compounds in bioassays.

Implementation: How I get the heatmap is that I get the bioassays from PubChem using the quary, dihydroorotate+dehydrogenase+and+Malaria, and function, find.assay.id. Then I use only the first 50 assays out of the 99 assays I got by using a loop. Here we have two heatmaps to generate: one is with active compounds and the other one is inactive. With the OUTCOME set to be “Active” or “Inactive”, I can filter the compounds. After getting compounds, use function, get.cid, to get the smiles of the compounds. Use function, parse.smiles, to parse the smiles. The statement, fp.sim.matrix(unlist(cmp.fp)), convert cmp.fp into a matrix and function, get.fingerprint, get the fingerprints from smiles. (Here we can see the bit sting when implement in separate line). Finally, we set the distance of the matrix and cluster the compounds. The heatmap produced is showing the distribution of similarity of the compounds in the dataset with Tanimoto coefficient as a measurement. The compounds displayed in columns and rows are showing in a hierarchy way which we called dendrogram and we can also see some similarity relation by the dendrogram.

Result: As we can see from the heat map (Fig 1), the active compounds in the dataset have high similarity (as there is darker yellow color represented similarity around 0.6). And the diagonal line is showed in white color as we expected because the diagonal line is the intersection of the same items. For the inactive compounds(Fig 2), there are only 3 compounds and showing low similarity.

Fig 1

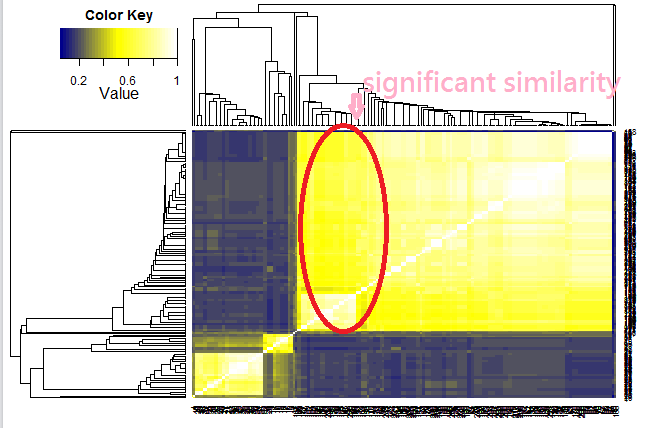


Fig 2

